基于随机子空间的多标签类属特征提取算法 *

张 晶，李 裕†，李培培 1

(合肥工业大学 计算机与信息学院, 合肥 230009)

摘 要：目前多标签学习已广泛应用到很多场景中，在此类学习问题中，一个样本往往可以同时拥有多个类别标签。由于类别标签可能带有的特有属性（即类属属性）将更有助于标签分类，所以已经出现了一些基于类属属性的多标签学习算法。针对类属属性构造会导致属性空间存在冗余的问题，本文提出了一种多标签类属特征提取算法 LIFT_RSM。该方法基于类属属性空间通过综合利用随机子空间模型及成对约束降维思想提取有效的特征信息，以达到提升分类性能的目的。在多个数据集上的实验结果表明：与若干经典的多标签算法相比，提出的 LIFT_RSM 算法能得到更好的分类效果。

关键词：多标签学习；成对约束；特征提取；随机子空间

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.08.0714

Multi-label label-specific feature extraction algorithm based on random subspace

Zhang Jing, Li Yu†, Li Peipei

(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: Multi-label learning has been widely used in many application scenarios right now. In this kind of learning problem, each instance is simultaneously assigned with more than one class label. Since different class labels might have their own unique characteristics (i. e. , label-specific feature) which would be more useful for label classification, so some multi-label learning approaches based on label-specific features had already been proposed. Therefore, aiming at the problem that redundant feature space caused by label-specific feature construction, a multi-label label-specific feature extraction algorithm named LIFT_RSM is proposed, which can improve the performance of classification by comprehensively using random subspace method and the thought of pair-wise constraint dimensionality reduction to extract effective feature information in label-specific feature space. The experimental results on several datasets show that the proposed algorithm can achieve better classification results compared with several classical multi-label algorithms.

Key Words: multi-label learning; pair-wise constraints; feature extraction; random subspace

0 引言

随着信息技术的发展，多标签学习[1-3]已逐渐成为数据挖掘领域的研究热点之一，得到了广泛的关注和研究。不同于传统单标签数据，在多标签数据中每个样本可同时隶属于多个标签，使得此类数据往往不再具有唯一语义。由于多标签数据的多义性特点，使得多标签学习在实际生活中可以广泛运用到许多应用场景中，并在如文本分类、音乐情感分类、语义场景分类、生物信息学及其他领域内取得了较好效果。

多标签学习问题可形式化地描述如下：给定 $X = R^d$ 代表 d 维样本空间，$L = \{l_1, l_2, \cdots, l_q \mid l \in \{0,1\}\}$ 表示包含 q 个标签的标签集合，其主要任务就是通过从训练集 $T = \{(x_i, Y_i) | i = 1, \cdots, p, x_i \in X, Y_i \subseteq L\}$ 中学习得到分类函数 $f : X \to Y$，将任意未知样本 $x \in R^d$ 映射到相应的标签集合 $L_x \subset L$。由于标签集合 L 中标签间的关系并不假定为互斥的，所以使得单标签学习框架不再适用于此类数据。

正因如此，经过近些年来许多学者的不断研究，一系列多标签算法被先后提出。总结目前已有的算法，其主要构造思路大致可分为以下三种：问题转换、算法适应和集成方法。问题转换方法[4-6]通过改造数据将多标签问题转换为若干个单标签问题，再利用成熟的单标签方法处理转换后的问题。此类方法虽然简单易行且不受特定算法的限制，但由于忽略了标签间的

基金项目：国家自然科学基金项目（61503112，61673152）；国家重点基础研究发展计划（973）项目（2016YFC0801406）；中央高校基本科研业务费专项资金（JZ2017HGBZ0930）

作者简介：张晶（1976-），女，安徽合肥人，副教授，博士，主要研究方向为人工智能、数据挖掘；李裕（1989-），男（通信作者），江西南昌人，硕士研究生，主要研究方向为数据挖掘（liyutty@163.com）；李培培（1982-），女，讲师，博士，主要研究方向为数据流分类、概念漂移检测方法、短文本分类、半监督学习、集成学习．

关联信息，会在一定程度上影响学习效果。算法适应方法[7-9]则直接扩展改进传统的单标签学习算法，增强其适用性和泛化能力，使之能适应多标签数据的处理。集成方法[10-11]通常将问题转换方法和算法适应方法结合起来处理多标签学习问题，以便取得更优的学习效果。

在处理多标签数据时，上述方法采用了一个相同的策略：即使用同一特征集合预测所有的类别标签。尽管此策略在多标签研究领域内取得了不错的效果，但其并非最优选择。由于每个标签可能具有独有的特征属性(即类属属性)，同时它们也是与标签最相关的属性，对相应标签具有更强的判别能力。基于此观点，Zhang 等提出了基于类属属性的 LIFT(multi-label learning with Label specific FeaTures)算法[12]。与已有策略不同，LIFT 算法借助类属属性确定未知样本的标签集合，然而其在类属属性的构造过程中，未充分考虑样本间的相关性，会导致类属属性维度增加，使得类属属性空间中存在冗余信息。

针对上述问题，本文综合利用随机子空间及成对约束降维的思想，提出了一种基于随机子空间的多标签类属特征提取算法，记为 LIFT-RSM。对于各个类属属性空间，该方法首先利用随机子空间思想将原始特征空间划分为多个部分；其次，在各个部分中利用近邻关系和成对约束获取相应权值矩阵；然后融合各权值矩阵并依此设计目标函数；最终通过矩阵的广义特征值分解学习得到变换矩阵，并以此构建对应的低维特征空间。实验结果表明，该算法取得了较好的分类效果，验证了算法的有效性。

1 相关工作

1.1 成对约束

在许多应用领域中，除样本的类别标记外，一些其他形式的背景知识也可以用作监督信息，其中就包括成对约束(*pair-wise constraints*)信息。成对约束是指某两个样本间的一种关系。相比于类别标记，成对约束适用范围更为广泛更为一般化，其不关注样本的具体类别，仅关心两个样本是否属于同一类别，因而更易获取。而且根据类别标记信息可以相对容易地获取等价的成对约束信息，反之则不然，因此成对约束比类别标记更具普遍意义。

成对约束通常可分为正约束(*must-link*, *ML*)和负约束(*cannot-link*, *CL*)两种[13]，其中正约束是指两个样本隶属于同一类别；相反地，负约束则要求两个样本属于不同类别。具体而言，对于给定的样本集合 $X=\left[x_1,x_2,\cdots,x_n\right]$，可将其中所有正约束的集合构成正约束集，形式化地表示为 $M=\{(x_i,x_j)|x_i,x_j\text{属于同一类}\}$；相应的，负约束集为所有负约束的集合，记为 $C=\{(x_p,x_q)|x_p,x_q\text{属于不同的类}\}$。

1.2 随机子空间

随机子空间是由 Ho[14-15]提出的一种有效的基于特征划分的集成学习方法，最初用于克服决策树分类器中的过学习问题。其基本思想是从原始特征空间中随机选取不同的特征子集并依此构建特征子空间，然后利用各特征子空间构造相应的子分类器，最后将通过不同子分类器学习得到的分类结果按照一定的组合规则进行融合集成，得到最终的学习决策。在特征随机选取过程中，不但能够更充分地利用原始特征信息、减少数据冗余，同时还能有效避免小样本问题。但由于特征选取的随机性，无法保证所选特征都包含有效判别信息，导致基分类器的准确性难以保证。

2 基于随机子空间的多标签类属特征提取算法

2.1 类属属性空间构建

LIFT 算法构建类属属性空间时需要考察各个标签下属性空间的内在性质。具体而言，对于任意标签 $l_k\in L$，可将训练集划分为正类样本集合 P_k 和负类样本集合 N_k 两部分，分别表示为

$$P_k=\left\{x_i|\left(x_i,Y_i\right)\in T,l_k\in Y_i\right\} \tag{1}$$

$$N_k=\left\{x_i|\left(x_i,Y_i\right)\in T,l_k\notin Y_i\right\} \tag{2}$$

由此可知，P_k 是由具有 l_k 标签的样本组成的集合；相反地，N_k 则由未被 l_k 标记的样本构成。

在文献[12]中，利用 *k-means* 算法分别对上述两个集合进行聚类分析。在此，可将集合 P_k 划分为 m_k^+ 个簇，其聚类中心记为 $\{p_1^k,p_2^k,\cdots,p_{m_k^+}^k\}$。相应地，集合 N_k 将被划分成 m_k^- 个簇，对应的聚类中心为 $\{n_1^k,n_2^k,\cdots,n_{m_k^-}^k\}$。文献[12]给予 P_k 和 N_k 的聚类信息相同的权重，因而将聚类中心的数目设为相等，即 $m_k^+=m_k^-=m_k$。具体来说，集合 P_k 和 N_k 的聚类数目将由以下公式确定：

$$m_k=\left\lceil \gamma\bullet min\left(\left|P_k\right|,\left|N_k\right|\right)\right\rceil \tag{3}$$

其中：$|\bullet|$ 表示集合的基数，$\gamma\in\left[0,1\right]$ 是控制聚类数目的参数。

由聚类的性质可知，上述两组聚类中心能够分别刻画对应集合的内在结构。因此，在此基础上，类属属性可按照如下公式进行定义：

$$\varphi_k\left(x_i\right)=\left[d\left(x_i,p_1^k\right),\cdots,d\left(x_i,p_{m_k}^k\right),d\left(x_i,n_1^k\right),\cdots,d\left(x_i,n_{m_k}^k\right)\right] \tag{4}$$

其中：$d\left(\bullet,\bullet\right)$ 返回两样本间的距离，在文献[12]中采用欧氏距离。

2.2 基于随机子空间的特征提取

2.2.1 随机子空间划分及融合

利用上文构建的类属属性空间，在原始 D 维空间中随机选取 P 个特征 $\left(P<D\right)$ 构建 T 个不同的 P 维子空间集合，记为 $F=\left\{F_1,F_2,\cdots,F_T\right\}$。其中，任一子空间 F_t 均为由 P 维样本 $f_i^t\in R^P$ 构成的空间，即 $F_t=\{f_1^t,f_2^t,\cdots,f_n^t\}$。为了清晰地描述子空间中样本的近邻关系，在此，利用距离均值来自适应的确定样本的近邻数。具体而言，就是在任意子空间 F_t 中，样本间的

近邻关系依据样本 f_i^t 与所有样本的距离均值 M_i^t 进行界定，即 $M_i^t=(\sum_{j=1}^{N}d_{ij}^t)/N$。当样本 f_i^t 与 f_j^t 之间的距离 d_{ij}^t 小于 M_i^t 时，将 f_i^t 视为 f_j^t 的近邻，否则二者间不存在近邻关系，如此不同样本的近邻数 k_i^t 一般是不相等的。

针对任意子空间 F_t，构建相应的自适应近邻图 G_t^N、非邻近图 G_t^F 及类间邻近图 G_t^B。具体来说，就是以图中的节点表示具体样本，利用图中的边来反映样本间的邻近关系。根据上述图关系分别定义各个样本与相应近邻样本的权重矩阵 $S_t^N=[S_{ij}^{t,N}]$、与相应非近邻样本的权重矩阵 $S_t^F=[S_{ij}^{t,F}]$、与相应类间邻近样本的权重矩阵 $S_t^B=[S_{ij}^{t,B}]$，上述矩阵的权重分别定义如下：

$$S_{ij}^{t,N}=\begin{cases}1, \ if \ d_{ij}^t<M_j^t \ or \ d_{ij}^t<M_i^t\\0, \ else\end{cases} \tag{5}$$

$$S_{ij}^{t,F}=\begin{cases}1, \ if \ d_{ij}^t\ge M_j^t \ and \ d_{ij}^t\ge M_i^t\\0, \ else\end{cases} \tag{6}$$

$$S_{ij}^{t,B}=\begin{cases}1, \ if \ \left(f_i^t,f_j^t\right)\in C \ and \ d_{ij}^t<F_i^t\\0, \ else\end{cases} \tag{7}$$

其中：d_{ij}^t 为样本间的欧氏距离，M_i^t 为样本 f_i^t 与所有样本距离的均值，F_i^t 表示样本 f_i^t 与其同类相距最远样本间的距离值。

为了能够更有效地利用子空间信息反映数据的真实分布情况，降低特征随机选取造成的不确定性。在此，分别融合已构建的 T 个自适应近邻图、T 个非邻近图及 T 个类间邻近图，得到相应的混合近邻图 G^N、混合非邻近图 G^F 及混合类间邻近图 G^B，并依据上述混合图关系构建对应的权值矩阵 S^{rsn}、S^{rsf} 和 S^{rsb}。以上混合图的权值矩阵均可借助各个子空间中相应权重矩阵进行线性重建得到。具体而言，可将它们之间的关系分别定义如下：

$$S_{ij}^{rsn}=\frac{1}{T}\sum_{t=1}^{T}S_{ij}^{t,N},S_{ij}^{rsf}=\frac{1}{T}\sum_{t=1}^{T}S_{ij}^{t,F},S_{ij}^{rsb}=\frac{1}{T}\sum_{t=1}^{T}S_{ij}^{t,B} \tag{8}$$

其中：S_{ij}^{rsn}、S_{ij}^{rsf}、S_{ij}^{rsb} 分别表示权值矩阵 S^{rsn}、S^{rsf} 及 S^{rsb} 中的权值。从上式可看出，混合图中的权值可由 T 个子空间中对应权重取均值获得。

2.2.2 设计目标函数

对于正约束关系 *ML*，为了能够有效保持类内整体的紧致性，本文将选取样本对应的全部同类样本用于构建权值矩阵 $S^m=[S_{ij}^m]$。因此，可以根据正约束集合 M 构造类内散布矩阵 Q_m 用于描述类内紧凑程度，定义如下：

$$\begin{aligned}Q_m&=\sum_{\substack{ij\\(x_i,x_j)\in M\ or\ (x_j,x_i)\in M}}\left(w^Tx_i-w^Tx_j\right)^2\\&=2w^TX\left(D^m-S^m\right)X^Tw\\&=2w^TXL^mX^Tw\end{aligned} \tag{9}$$

$$S_{ij}^m=\begin{cases}1, \ if \ \left(x_i,x_j\right)\in M \ or \ \left(x_j,x_i\right)\in M\\0, \ else\end{cases} \tag{10}$$

其中：S^m 为对称矩阵，D^m 为对角矩阵，其对角线上的元素是矩阵 S^m 中相应的列(或行)和，即 $D_{ii}^m=\sum_j S_{ij}^m$。$L^m=D^m-S^m$ 为拉普拉斯矩阵，是一个对称的半正定矩阵。

对于负约束关系 *CL*，为了能够充分反映样本间的差异性，在这里，本文利用混合类间邻近图 G^B 对原始负约束集合 C 进行调整，构造新的负约束集合。并以相应的权值矩阵 S^{rsb} 为基础，构建可以刻画类间离散程度的类间混合散布矩阵 Q_{rsb}，定义如下：

$$\begin{aligned}Q_{rsb}&=w^TX\left[\left(D^{col}+D^{row}\right)-\left(S^{rsb}+\left(S^{rsb}\right)^T\right)\right]X^Tw\\&=w^TX\left(\bar{D}^{rsb}-\bar{S}^{rsb}\right)X^Tw\\&=w^TXL^{rsb}X^Tw\end{aligned} \tag{11}$$

其中：S^{rsb} 为非对称矩阵，D^{col} 和 D^{row} 均为对角矩阵，即 $D_{ii}^{col}=\sum_j S_{ij}^{rsb},D_{jj}^{row}=\sum_i S_{ij}^{rsb}$，$L^{rsb}=\overline{D}^{rsb}-\overline{S}^{rsb}$。

到目前为止仅考虑了与成对约束有关的信息，尚未涉及样本集所包含的潜在信息。在此，为了能够充分利用样本间的邻近信息，可以基于流形假设[16]将样本间的近邻关系作为局部结构信息导入降维过程中。一方面，希望在原始空间中相互靠近的样本其投影在低维空间中也是互相靠近的。因此，根据混合近邻图 G^N 的权值矩阵 S^{rsn}，可以构建混合邻近散布矩阵 Q_{rsn} 用于描述近邻点之间的紧密程度，具体定义如下：

$$\begin{aligned}Q_{rsn}&=\sum_{ij}\left(w^Tx_i-w^Tx_j\right)^2S_{ij}^{rsn}\\&=2w^TX\left(D^{rsn}-S^{rsn}\right)X^Tw\\&=2w^TXL^{rsn}X^Tw\end{aligned} \tag{12}$$

其中：D^{rsn} 为对角矩阵，$D_{ii}^{rsn}=\sum_j S_{ij}^{rsn}$，$L^{rsn}=D^{rsn}-S^{rsn}$。

另一方面，对于非近邻样本，期望其在低维空间中的投影点能够尽可能的散开。基于此，利用混合非邻近图 G^F 的权值矩阵 S^{rsf} 定义了下式用于度量非近邻样本间的散开程度：

$$\begin{aligned}Q_{rsf}&=\sum_{ij}\left(w^Tx_i-w^Tx_j\right)^2S_{ij}^{rsf}\\&=2w^TX\left(D^{rsf}-S^{rsf}\right)X^Tw\\&=2w^TXL^{rsf}X^Tw\end{aligned} \tag{13}$$

其中：Q_{rsf} 表示混合非邻近散布矩阵，D^{rsf} 代表对角矩阵，$D_{ii}^{rsf}=\sum_j S_{ij}^{rsf}$，$L^{rsf}=D^{rsf}-S^{rsf}$。

基于上述准备，在设计目标转换向量 w^* 时，应该以成对约束信息为指导，同时充分利用样本间的近邻关系。因此，最终目标转换向量可以通过定义如下函数得到：

$$w^*=\underset{w}{argmax}\frac{Q_{rsb}+\alpha Q_{rsf}}{Q_m+\beta Q_{rsn}}=\underset{w}{argmax}\frac{w^TX\left(L^{rsb}+\alpha L^{rsf}\right)X^Tw}{w^TX\left(L^m+\beta L^{rsn}\right)X^Tw} \tag{14}$$

其中：α 和 β 为常系数，分别用于调节 Q_{rsf} 和 Q_{rsn} 的贡献度。如果 $X(L^m+\beta L^{rsn})X^T$ 为非奇异的，那么可以使用拉格朗日方法变换上式，将上式的求解问题转换为如下等式求解最大广义特征值对应特征向量的问题：

$$X\left(L^{rsb}+\alpha L^{rsf}\right)X^T w=\lambda X\left(L^m+\beta L^{rsn}\right)X^T w \tag{15}$$

根据阈值参数 thr $(0\le thr\le 1)$，通过 $\sum_{i=1}^{d}\lambda_i \ge thr\times\sum_{i=1}^{D}\lambda_i$ 确定最终维度 d，并选取前 d 个最大非零特征值的对应特征向量构成变换矩阵 W。

2.3　算法描述

本节将随机子空间思想引入类属空间，充分利用成对约束信息及样本的近邻关系，提出了一种基于随机子空间的多标签类属特征提取算法。以下完整地展示了从类属属性构建、子空间划分融合、特征提取、分类模型训练至未知样本预测的全部流程，其详细操作过程可以总结如下：

输入：训练集 X,聚类个数控制参数 γ,随机子空间个数 T,特征子空间维度 P,贡献度控制参数 α 和 β,阈值参数 thr,未标记样本 x'；

输出：预测标签集合 Y'.

① 对于每一种类标签 l_k，重复步骤 2~15；

② 根据式(1)~(2)，利用训练集 X 构建样本集 P_k 和 N_k；

③ 在 P_k 和 N_k 上，用 k-means 算法进行聚类分析，聚类个数 m_k 根据式(3)获得；

④ 根据式(4)构建原始类属属性空间 $L_k=tr_k\cup ts_k$，其中 $tr_k=\{\varphi_k(x_i),\forall x_i\in X\}$，$ts_k=\{\varphi_k(x')\}$；

⑤ 对 L_k 进行中心化，得到类属属性空间 $C_k=Tr_k\cup Ts_k$；

⑥ 在 Tr_k 中随机选取 P 维特征构成子空间 F_t；

⑦ 在 F_t 上构造近邻图 G_t^N、非邻近图 G_t^F 和类间邻近图 G_t^B，并根据式(5)~(7)计算对应权重矩阵；

⑧ 返回步骤 6，如此循环 T 次；

⑨ 利用各个子空间中的图关系构建混合图，根据式(8)计算各混合图权值矩阵；

⑩ 根据式(9)~(13),分别构建散布矩阵 Q_m、Q_{rsb}、Q_{rsn} 和 Q_{rsf}；

⑪ 确定权值 α 和 β，构造目标转换函数如式(14)所示；

⑫ 根据 thr 确定维度 d，求解式(15)得到变换矩阵 W_k，通过 $My_k=W_k^T Tr_k$ 得到降维后的类属属性空间，即映射 $\rho_k(a_i),\forall a_i\in Tr_k$；

⑬ 以映射 $\rho_k(a_i)$ 为基础构建相应二分类训练集 T_k^*；

⑭ 基于 T_k^* 使用二分类学习算法得到相应的分类模型 $f_k: My_k\to R$；

⑮ 预测的标签集合 $Y'=\{l_k\,|\,f_k(\rho_k(t))>0,1\le k\le q,t\in Ts_k\}$。

LIFT-RSM 算法首先为每个类标签构建类属属性空间并中心化（步骤 2~5）；然后利用随机子空间思想划分原始类属空间，融合各子空间的近邻关系后，借助成对约束信息对原始类属空间进行降维(步骤 6~12)；接着在降维后的类属属性空间中训练二分类模型(步骤 13~14)；最后对未知样本进行预测(步骤 15)。

3　实验分析

3.1　数据集

本文采用 Scene、Emotions、Slashdot、Flags 和 Image 等 5 种不同的公开多标签数据集，对提出的基于随机子空间的多标签类属特征提取算法进行实验验证，上述数据集的具体统计信息如表 1 所示。由于所选的数据集涵盖了音乐、图像、文本等不同应用领域，而且标签性质各不相同，因而具有较强的概括性。

表 1　数据集信息

数据集	|S|	*dim*(*S*)	*L*(*S*)	*LCard*(*S*)	*LDen*(*S*)	URL
Image	2000	294	5	1.236	0.247	URL2
Scene	2407	294	6	1.074	0.179	URL1
Emotions	593	72	6	1.869	0.311	URL1
Flags	194	19	7	3.392	0.485	URL1
Slashdot	3782	1079	22	1.180	0.054	URL3

注：URL1:http://mulan.sourceforge.net/datasets-mlc.html

URL2:http://cse.seu.edu.cn/PersonalPage/zhangml/index.htm

URL3:http://computer.njnu.edu.cn/Lab/LABIC/LABIC_software.html

在表 1 中：$|S|$ 表示样本个数；$dim(S)$ 表示属性个数；$L(S)$ 表示标签个数；$LCard(S)$ 表示标签基数，为样本具有的平均相关标签个数；$LDen(S)$ 表示标签密度，为由标签个数归一化的标签基数。

3.2　实验设置

3.2.1 评估指标

在多标签学习中，由于每个样本可以同时隶属于多个标签，所以通常检验多标签算法的有效性与检验单标签算法相比更加复杂。在传统单标签算法中广泛应用的评价指标如准确率、查全率、精度等已不再适用于多标签问题，为此需要引入专门的多标签评价指标来验证算法的有效性。目前，多标签评价指标主要从样本和标签两个角度度量算法的性能，可大致分为两类[1]：即基于样本的指标[17]和基于标签的指标[18]。在本文实验中，选取以下 5 项评价指标来综合评估提出算法的性能，其中包括 1 个基于样本的指标：汉明损失(Hamming Loss, *HL*)及 4 个基于标签排序的指标：1-错误率(One-Error,*OE*)、排序损失(RankingLoss, *RL*)、覆盖率(Coverage,*CV*)、平均精度(Average Precision, *AP*)。

上述五种指标分别从不同角度评价算法性能的优劣，并直接反映在指标数值的大小上。其中，平均精度在值越大的时候算法性能越好，当其值为 1 时，性能达到最优；余下 4 个评价指标，取值越小表示算法性能越好，所以当值为 0 时，性能最

好，反之为 1 时最差。有关上述评价指标的详细介绍具体可参照文献[1]，在此不再赘述。

3.2.2 对比算法

本文选取 5 种经典的多标签学习方法用作对比算法，分别与本文提出的 LIFT-RSM 算法进行对比及分析。这 5 种算法包括：基于 k 近邻的 ML-kNN 算法[7]、LIFT 算法[12]、多标签维度约减算法 MDDM[19]、MLNB[20]和 MLSI[21]。实验中，对于 LIFT 和 LIFT-RSM 算法，将参数 γ 以 0.1 为步长在[0,1]区间内进行调节，并最终设定 $\gamma = 0.2$；对算法 MLNB、MLSI，设置保持原数据 98%的信息量；对于 MDDM 及 ML-kNN 算法，根据相应文献的建议选取默认的参数配置。

如无特别说明，在 LIFT_RSM 算法中控制贡献度的参数 α 和 β 均设置为 0.05，随机子空间个数 T 设置为 10，特征子空间维度 P 设置为 20，若原始空间的维度 d 小于 20 时，则将 P 设置为 $\lfloor 0.3*d \rfloor$，在降维过程中保留原类属属性 95%的信息量。除 MLNB 算法外，使用线性核 LIBSVM 作为其余所有算法的基分类器。本文所有实验在内存为 4GB 及 2.50GHz 处理器的主机上完成，操作系统为 64 位 Windows 7，并选取 MATLAB 2014a 作为开发平台。

3.3 结果分析

对于 Image、Scene、Flags 数据集，本文从数据集中随机抽取 80%的样本作为训练集，余下 20%的样本组成测试集，抽样过程重复 50 次并记录 50 次实验的均值。余下数据集使用原始训练集和测试集，重复实验 50 次并记录 50 次实验的均值。

表 2-6 分别记录了各个算法在 5 种数据集上的实验结果，实验结果采用均值形式表示。其中，对于各项评价指标，符号↓（↑）表示该项评价指标值越小（越大）算法性能越优。此外，各项评价指标的最优值以下划线方式标出。

表 2 数据集 Scene 分类性能比较

算法	HL↓	OE↓	CV↓	RL↓	AP↑
MLSI	0.1085	0.2684	0.5645	0.0955	0.8376
ML-kNN	0.8798	0.2279	0.4780	0.0782	0.8641
MDDM	0.1065	0.2605	0.5194	0.0877	0.8454
MLNB	0.8846	0.2797	0.5729	0.0962	0.8331
LIFT	0.0770	0.1903	0.3839	0.0615	0.8878
LIFT_RSM	0.0762	0.1818	0.3882	0.0606	0.8918

表 3 数据集 Flags 分类性能比较

算法	HL↓	OE↓	CV↓	RL↓	AP↑
MLSI	0.3422	0.2606	3.9308	0.2410	0.7909
ML-kNN	0.6872	0.2376	3.9846	0.2426	0.7934
MDDM	0.3379	0.2531	3.8749	0.2318	0.7960
MLNB	0.6215	0.2813	4.3803	0.3300	0.7400
LIFT	0.3382	0.2390	3.8144	0.2330	0.8001
LIFT_RSM	0.3168	0.2274	3.9046	0.2300	0.8012

表 4 数据集 Emotions 分类性能比较

算法	HL↓	OE↓	CV↓	RL↓	AP↑
MLSI	0.2946	0.4307	2.6733	0.3236	0.6723
ML-kNN	0.8762	0.4059	2.4901	0.2829	0.6938
MDDM	0.2698	0.3762	2.2475	0.2367	0.7307
MLNB	0.8985	0.4257	2.3762	0.2706	0.7051
LIFT	0.2700	0.3483	2.4110	0.2506	0.7302
LIFT_RSM	0.2607	0.3478	2.3144	0.2387	0.7352

表 5 数据集 Image 分类性能比较

算法	HL↓	OE↓	CV↓	RL↓	AP↑
MLSI	0.1904	0.3453	1.0186	0.1867	0.7770
ML-kNN	0.8820	0.3224	0.9781	0.1774	0.7893
MDDM	0.2003	0.3732	1.0633	0.1964	0.7591
MLNB	0.8572	0.4143	1.2605	0.2462	0.7223
LIFT	0.1535	0.2654	0.8305	0.1398	0.8286
LIFT_RSM	0.1512	0.2566	0.8347	0.1406	0.8313

表 6 数据集 Slashdot 分类性能比较

算法	HL↓	OE↓	CV↓	RL↓	AP↑
MLSI	0.0549	0.4865	3.2267	0.1271	0.6253
ML-kNN	0.0528	0.6642	4.2624	0.1785	0.4775
MDDM	0.0470	0.6484	4.2201	0.1793	0.4952
MLNB	0.9722	0.5577	5.5558	0.1470	0.2379
LIFT	0.0400	0.4163	2.4545	0.0968	0.6813
LIFT_RSM	0.0397	0.4096	2.4202	0.0948	0.6871

观察表 2~6 中的结果可以看出，本文提出的基于随机子空间的多标签类属特征提取算法 LIFT_RSM 取得了较好的分类效果。对于 Emotions 和 Image 数据集，除了覆盖率和排序损失，LIFT_RSM 算法在余下的 3 个评价指标上均优于其他对比算法。对于 Scene 和 Flags 数据集，除覆盖率外，LIFT_RSM 算法的余下 4 项指标均优于对比算法。对于 Slashdot 数据集，LIFT_RSM 算法在 HL、OE、CV、RL 和 AP 等 5 项指标上的结果分别为 0.0397、0.4096、2.4202、0.0948 及 0.6871，相比于原始 LIFT 算法均有不同程度的提升，与其他算法相比效果提升更为显著。值得注意的是，部分数据集中提出算法在部分指标上表现略差。经过分析，发现其主要原因是相关数据集的标签密度较大，同时拥有多个标签的实例较多，使得各个实际类别中边缘样本及噪声样本增多，致使特征提取性能受到影响，导致分类效果不理想。

以 Flags、Scene、Emotions 及 Image 数据集为例，分别统计上述数据集用 LIFT 和 LIFT_RSM 算法学习后得到的类属属性维度，具体对比情况如图 1~4 所示。

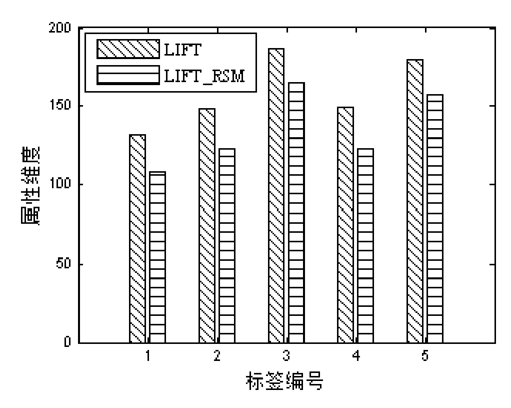


图 1 数据集 Image 类属属性维度比较

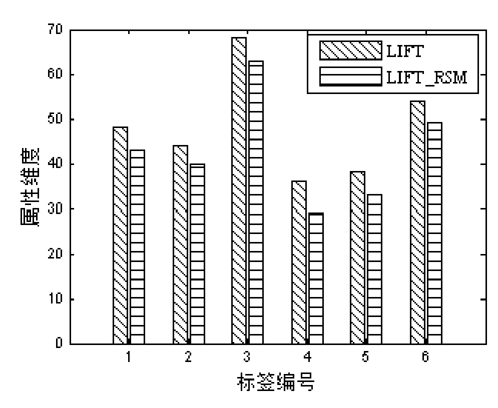


图 2 数据集 Emotions 类属属性维度比较

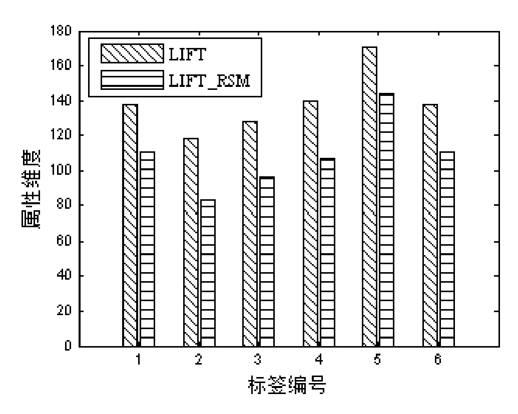


图 3 数据集 Scene 类属属性维度比较

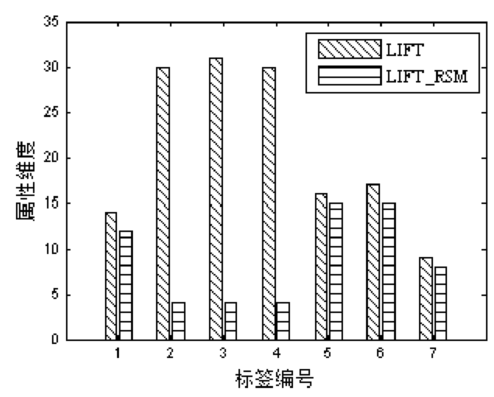


图 4 数据集 Flags 类属属性维度比较

从图 1~4 中可以发现，针对不同的数据集，在其对应的所有类别标签上，LIFT_RSM 算法得到的最终类属属性维度均不同程度低于 LIFT 算法的类属属性维度。其中，以 Scene 数据集为例，在其包含的 6 个标签上，LIFT 算法学习得到的原始类属属性维度分别为 138、118、128、139、171、138，而使用 LIFT_RSM 算法进行学习后对应属性维度分别下降至 110、83、96、107、144、110，由此可见由 LIFT_RSM 算法学习得到的类属属性维度的确能够有一定程度的降低。虽然 LIFT_RSM 算法在部分评价指标上略低于对比算法，但总体而言，LIFT_RSM 算法仍然能够获得较好的学习分类性能。

LIFT_RSM 算法通过融合各个随机子空间中样本的近邻关系，可以更精确的表示样本间的相关性，因而可以有效解决多标签数据分类问题。综上所述，本文提出的 LIFT_RSM 算法在综合性能上总体优于其他对比算法，提高了分类器的性能，并取得了较好的效果。

4 结束语

不同与以往多标签学习算法，LIFT 算法着重考察属性空间操作对多标签学习性能的影响。本文以 LIFT 算法为基础，利用随机子空间模型划分原始类属空间，在融合各个子空间中近邻关系后，借助成对约束信息指导降维的思想，提出了一种基于随机子空间的多标签类属特征提取算法。一系列实验结果表明，提出算法整体上优于其他经典算法，符合预期目标，验证了该算法的有效性。在今后的研究中，可以将标签间的相关性融入到类属属性特征提取中，以进一步提升多标签算法的学习性能。此外，目前该算法的参数个数相对较多，寻找有效的自适应方法减少所需参数数目也是未来的研究工作之一。

参考文献：

[1] Zhang M L, Zhou Z H. A Review on multi-label learning algorithms [J]. IEEE Trans on Knowledge & Data Engineering, 2014, 26 (8): 1819-1837.

[2] Tsoumakas G, Katakis I. Multi-label classification: an overview [J]. International Journal of Data Warehousing & Mining, 2009, 3 (3): 1-13.

[3] Zhou Z H, Zhang M L. Multi-label learning [M]// Encyclopedia of Machine Learning and Data Mining, Berlin: Springer, 2017, 875-881.

[4] Zhang M L, Zhang K. Multi-label learning by exploiting label dependency [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 999-1008.

[5] Fürnkranz J, Hüllermeier E, Mencía E L, et al. Multilabel classification via calibrated label ranking [J]. Machine Learning, 2014, 73 (2): 133-153.

[6] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification. [J]. Pattern Recognition, 2004, 37 (9): 1757-1771.

[7] Zhang M L, Zhou Z H. M L-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40 (7): 2038-2048.

[8] Zhang M L, Zhou Z H. Multilabel neural networks with applications to functional genomics and text categorization [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18 (10): 1338-1351.

[9] Schapire R E, Singer Y. BoosTexter: A Boosting-based System for Text

Categorization [J]. Machine Learning, 2000, 39 (2-3): 135-168.

[10] Tsoumakas G, Vlahavas I. Random k-labelsets: an ensemble method for multilabel classification [C]// Proc of the 18th European Conference on Machine Learning. Berlin: Springer, 2007: 406-417

[11] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification [J]. Machine Learning, 2011, 85 (3): 254-269.

[12] Zhang M L, Wu L. Lift: Multi-label learning with label-specific features. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2015, 37 (1): 107-20.

[13] Klein D, Kamvar S D, Manning C D. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering [C]// Proc of the 19th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc. 2002: 307-314.

[14] Ho T K. Random decision forests [C]// Proc of International Conference on Document Analysis and Recognition. 2002: 278.

[15] Ho T K. The random subspace method for constructing decision forests [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 1998, 20 (8): 832-844.

[16] 刘建伟，刘媛，罗雄麟. 半监督学习方法 [J]. 计算机学报, 2015, 38 (8): 1592-1617.

[17] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification [C]// Lecture Notes in Computer Science. 2004, : 22-30.

[18] Yang Y. An evaluation of statistical approaches to text categorization [J]. Information Retrieval Journal, 1999, 1 (1): 69-90.

[19] Zhang Y, Zhou Z H. Multilabel dimensionality reduction via dependence maximization [J]. ACM Trans on Knowledge Discovery from Data, 2010, 4 (3): 1503-1505.

[20] Zhang M L, Peña J M, Robles V. Feature selection for multi-label naive Bayes classification [J]. Information Sciences, 2009, 179 (19): 3218-3229.

[21] Yu K, Yu S, Tresp V. Multi-label informed latent semantic indexing [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005: 258-265.